New Journal of Physics
The open access journal at the forefront of physics
Deutsche Physikalische Gesellschaft DPG | IOP Institute of Physics

The ten thousand Kims

New Journal of Physics
The open-access journal for physics

The ten thousand Kims

Seung Ki Baek[1], **Petter Minnhagen**[1,4] **and Beom Jun Kim**[2,3]
[1] Integrated Science Laboratory, Department of Physics, Umeå University, 901 87 Umeå, Sweden
[2] BK21 Physics Research Division and Department of Physics, Sungkyunkwan University, Suwon 440-746, Korea
[3] Asia Pacific Center for Theoretical Physics, Pohang 790-784, Korea
E-mail: Petter.Minnhagen@physics.umu.se

New Journal of Physics **13** (2011) 073036 (12pp)
Received 12 June 2011
Published 27 July 2011
Online at http://www.njp.org/
doi:10.1088/1367-2630/13/7/073036

Abstract. In Korean culture, the names of family members are recorded in special family books. This makes it possible to follow the distribution of Korean family names far back in history. It is shown here that these name distributions are well described by a simple null model, the random group formation (RGF) model. This model makes it possible to predict how the name distributions change and these predictions are shown to be borne out. In particular, the RGF model predicts that for married women entering a collection of family books in a certain year, the occurrence of the most common family name 'Kim' should be directly proportional to the total number of married women with the same proportionality constant for all the years. This prediction is also borne out to a high degree. We speculate that it reflects some inherent social stability in the Korean culture. In addition, we obtain an estimate of the total population of the Korean culture down to the year 500 AD, based on the RGF model, and find about ten thousand Kims.

[4] Author to whom any correspondence should be addressed.

Contents

1. Introduction

One's family name is very important in Korean culture. Abandoning one's family name is extremely unusual and considered dishonorable. A common metaphor is to pledge one's family name to a given promise. The importance of family names is also reflected by the fact that a married woman's name carries the family name of the family she comes from. In addition, the Confucian tradition has encouraged a family to record the genealogical tree in special books, which then records the women's family names flowing into the family book by marriage. The children inherit the name of the father. Some of the family books go back to more than 500 years ago.

The distribution of family names is often described in terms of the probability $P(k)$ to randomly pick a family name which occurs k times within the population. The frequency distributions $P(k)$ for family names are examples of broad 'fat-tailed' distributions that, at least crudely, can be described by power laws $P(k) \propto 1/k^{\gamma}$, as was studied earlier [1–3]. In particular, Korean family names have a very broad distribution with a γ close to 1 [4]. It is common practice to try to connect the approximate power-law form of family distributions to growth models of the total population [1–3, 5]. Such models usually yield power laws with approximately the Zipf's law exponent $\gamma = 2$, i.e. $P(k) \propto 1/k^2$. However, this does not describe the Korean family distribution, which has a much slower falling-off at large k. It has been suggested that this is because the rate of introducing new family names in Korea is very slow [3, 5]. In agreement with this, it was in [5] shown that a growth model with an introduction of family names which approaches zero can indeed yield a power law with $\gamma = 1$ instead of $\gamma = 2$.

In [6], it was shown that system-specific growth models are usually too restrictive to catch some of the essential characteristic features of frequency distributions. Examples are the dependence of γ on the size of the data set and the connection between the data set size and the size of the largest frequency. In the present paper, we reinvestigate the historical Korean family books from this particular perspective. We use a collection of Korean family books to estimate the change in frequency distribution of family names for the last 500 years. We then compare these changes with the predictions of the random group formation (RGF) model introduced in [6]. The RGF model assumes maximal mixing for each given data size and it is shown that the predictions from this model are borne out to a striking degree. In particular, it is found that the proportion of persons named 'Kim' is constant irrespective of all social changes, wars, earthquakes, famines, plagues, fertility variations, industrial revolution, etc and this constancy is also an inherent feature of the RGF model.

Table 1. Statistical quantities of the family books analyzed in this paper. Here, M is the total number of women entering the family by marriage at the specified period. N is the number of different family names that the women carried. Among these N different names, we find the one with the largest number of carriers and denote this number by $k_{\max}$.

Number	Period	M	N	$k_{\max}$
1	1510–1540	33	19	6
2	1540–1570	94	31	23
3	1570–1600	215	38	40
4	1600–1630	384	48	88
5	1630–1660	643	52	119
6	1660–1690	1039	59	230
7	1690–1720	1534	74	311
8	1720–1750	2313	82	407
9	1750–1780	3524	83	598
10	1780–1810	5640	96	1069
11	1810–1840	8499	110	1502
12	1840–1870	13 028	114	2256
13	1870–1900	19 559	129	3410
14	1900–1930	37 531	153	6888
15	1930–1960	79 935	163	15 735
16	1960–1990	47 554	162	9693

In section 2, we describe how we use the data from the Korean family books, and in section 3, we give a brief recapitulation of the RGF model. A comparison of the data with theoretical predictions is given in section 4, whereas section 5 contains the concluding remarks.

2. Korean family books

Our data are extracted from the ten Korean family books that were also analyzed in [7]. The data we extract in the present investigation are the total number M of married women, who were registered with marriage year into these ten books during specific 30-year periods between 1510 and 1990. For each period, the number of different family names N and the number of women having the most common family name (usually 'Kim'), $k_{\max}$, are also extracted. The results are given in table 1 which contains 16 historical windows. This data set is analyzed and compared with census data for the whole Korean population from the year 2000 (see [4]).

As in [7], we argue that the statistics of this collection of women's family names should bear a strong resemblance to the whole population during the same period in the following sense: suppose that out of the whole population of women who got married at a certain period, you have randomly selected M women. There is a certain chance that a chosen woman has a family name that occurs k times among the M picked women. Suppose that the probability distribution for the frequency of different family names within a group of M randomly selected women is $P_M(k)$. Then the M selected women on average have family names that occur M/N times since $\sum_{k=1}^{k_{\max}} k P_M(k) = \langle k \rangle = M/N$, where $k_{\max}$ is the most frequent name. The data for an entry in table 1 correspond to a single try of choosing M women out of the whole collection of women who got married during the period. Suppose that you have now instead picked M random persons out of the whole population. Then provided that the married women were really randomly distributed over the population, the result would be statistically the same: the probability distribution for the frequency of different family names within a group of M randomly selected persons would again be given by $P_M(k)$.

There are two important points to note in this context. The first is that $P_M(k)$ for M randomly selected persons does not have the same functional form as $P_{M_{\rm tot}}(k)$ for the complete population $M_{\rm tot}$ [6]. In other words, the family-name distribution depends on the size of M. The second is that $M_{\rm tot}(t)$ depends on time t in an unknown (but presumably rather nontrivial) way reflecting the history of the Korean people. There is, at least *a priori*, no obvious relation between the inflow of married women into the ten specific family-name books, on the one hand, and the total population of the Korean people on the other: a prosperous family could have a large inflow of married women even in periods when the total population decreases.

3. The random group formation model

The RGF model tries to catch the essential features of the group-size distribution when M objects are divided into N groups by assuming optimal mixing [6]. In the case of family names, the persons are the objects and the groups are formed by the persons carrying the same family name. The RGF model does not make any explicit assumption about what particular process is responsible for the creation of the groups. Instead it is assumed that, whatever this process might be, the result is that the optimal mixing condition is on average approximately fulfilled at all times. The optimal mixing condition corresponds to a maximum entropy condition for the group-size distribution $P_M(k)$. The appropriate maximum condition can be formulated in terms of a maximum mutual-information principle or equivalently as a minimum information-cost condition [6, 8]. The result is a distribution function of the explicit form

$$P_M(k) = A\frac{\exp(-bk)}{k^\gamma}, \tag{1}$$

where $P_M(k)$ obeys the following set of self-consistent equations:

$$\begin{cases} \displaystyle\sum_{k=k_0}^{M} A\frac{\mathrm{e}^{-bk}}{k^\gamma} = 1, \\ \displaystyle\sum_{k=k_0}^{M} Ak\frac{\mathrm{e}^{-bk}}{k^\gamma} = M/N, \end{cases}$$

where k_0 is the size of the smallest group. In the present investigation, this limit is always $k_0 = 1$, but in other applications it can be generalized to an arbitrary k_0. This means that the constants γ and b are interdependent through the relation:

$$\frac{\sum_{k=k_0}^{M}\frac{\mathrm{e}^{-bk}}{k^{\gamma-1}}}{\sum_{k=k_0}^{M}\frac{\mathrm{e}^{-bk}}{k^{\gamma}}} = \frac{M}{N}. \tag{2}$$

The self-consistency condition connects the entropy of $P_M(k)$ to the size of the largest group $k_{\max}$, through the relation

$$\langle k_{\max}\rangle = \frac{\sum_{k=k_c}^{M} kP(k)}{\sum_{k=k_c}^{M} P(k)}, \tag{3}$$

where $\langle k_{\max}\rangle$ is the average size of the largest group and the value of k_c is determined such that there is on average only a single group in the interval $[k_c, M]$, i.e. $\sum_{k=k_c}^{\infty} P(k) = 1/N$, which means that (3) gives the average $\langle k_{\max}\rangle$. When analyzing data we will approximate $k_{\max}$ for the data with the average size of the largest group $\langle k_{\max}\rangle$ obtained for the RGF model. This set of self-consistent equations yields a unique solution $P_M(k)$ of the RGF form for given values of M, N, $k_{\max}$ and k_0 [6]. Suppose you have a collection of M persons each carrying one out of N different family names which are distributed according to the frequency distribution $P_M(k)$. What is the corresponding frequency distribution $P_m(k)$ for m persons randomly picked from the original total M? In the case when the persons are picked randomly, $P_m(k)$ is given by the transformation

$$P_m(k) = \frac{\sum_{k'=k}\left(\frac{m}{M-m}\right)^{k}\left(\frac{M-m}{m}\right)^{k'}\binom{k'}{k}P_M(k')}{1-\sum_{k'=1}\left(\frac{M-m}{M}\right)^{k'}P_M(k')},$$

where $\binom{k'}{k}$ is the binomial coefficient. In the context of word frequency of books, this transformation is sometimes referred to as the random book transformation (RBT) [9, 10]. The crucial point is that, if you start with a certain $P_M(k)$ of the RGF form of (1), then all of γ, b and A will change with m: a random reduction of a data set is not scale-invariant with respect to the frequency distribution and, as a consequence, the power-law index γ changes [6, 9, 10]. Another characteristic feature of the transformation is that the number of persons in the largest family group is proportional to the size of the data set [6].

In order to predict the change in the RGF function under the data-size reduction, we first numerically calculate $\langle k\rangle_m = \sum_{k=1}^{m} kP_m(k)$ using the RBT transformation and then use the RGF self-consistent equations for the input values m, $n = m/\langle k\rangle_m$ and $k_{\max}(M)m/M$, to obtain the corresponding $P_m(k)$ of the form of (1). In this way, one obtains predictions for $P_m(k)$ starting from a known $P_M(k)$ when $m < M$. It is also possible to get predictions for an increase of the data set (i.e. $m > M$) in a similar way. However, in this paper, we will, when predicting the increase of a data set, resort to two approximate relations found in [6]

$$\gamma(M) - 1 \propto \frac{1}{\log M}, \tag{4}$$

describing the expected approach to the large-M limit together with the approximate relation $b \propto 1/M$.

Earlier attempts to explain family-name distributions have usually been connected to explicit assumption about the time evolution linked to the growth of the population [1–3, 5]. One may then ask how the time evolution enters the RGF model. The answer is that it only enters indirectly; the RGF model is in itself history-independent and at each time only depends on the instantaneous input parameters. However, one of the input parameters is the size of the data set M. Suppose that M is the total population; then this parameter is indeed history- and time-dependent. One might even suspect that it has a complicated time dependence $M(t)$ reflecting changes due to wars, earthquakes, famines, plagues, fertility variations, industrial revolution,

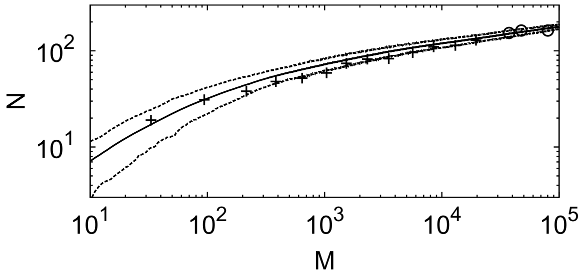

Figure 1. Comparison between the expected number of different family names and data. The full drawn curve is the number of different family names N as a function of women's names M, when M women are randomly chosen from all the women's names recorded in the ten books given in table 1 during the period 1900–1990. The curve is the average over 10^2 random choices, and the dotted lines show three standard deviations. The three open circles are the explicit data from table 1 for the three periods 1900–1930, 1930–1960 and 1960–1990. The crosses represent the remaining historical data which decrease monotonically with time. The data are consistent with a random drawing of persons from a *time-independent* distribution.

etc. The point is that the RGF model assumes that, whatever this actual historical time dependence might be, the resulting family-name frequency distribution to a good approximation on average is given by the maximal mixing condition, which only depends on the instantaneous value of $M(t)$.

A parallel example of this is provided by the word-frequency distribution of novels written by an author [6, 10]: no matter what the size of the novel or when it was written, to good approximation the word-frequency distribution for a novel of an author only depends on the number of words it contains [10]. This size dependence is to very good approximation given by the RGF model [6]. The fact that the word frequency of an author to good approximation only depends on the size of the text is equivalent to characterizing an author by a single, very large 'meta-book' from which the average frequency distribution from any text size written by the author can be obtained [10].

4. Analysis

In order to test the RGF model, we start with the three most recent entries in table 1 which span the time period 1900–1990. These three entries together contain $M_{1900\text{–}90} = 165\,020$ women each having one family name out of $N_{1900\text{–}90} = 194$. Out of this data set, we randomly select $M < M_{1900\text{–}90}$ women and calculate the average N for each M. The full curve in figure 1 shows the average $N(M)$. This random selection prediction based on the specific period 1900–1990 is compared with the data covering all the time periods from 1500 to 1990. If the women flowing into the entries by marriage are statistically equivalent to selecting random persons in the total population and if the population was static in time, then the agreement between the data and the random selection would be easy to understand. However, the latter assumption is of course not true: both the total population $M_{\rm tot}(t)$ and the number of different names $N_{\rm tot}(t)$ change in response to historical developments: from around the year 1500 to 2000, the population in

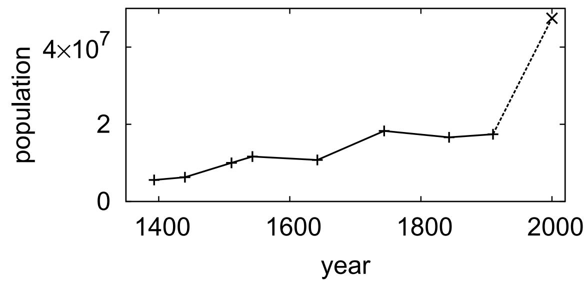

Figure 2. Historical estimates of the Korean population in [11] (crosses). The rightmost point indicates the census data in the year 2000.

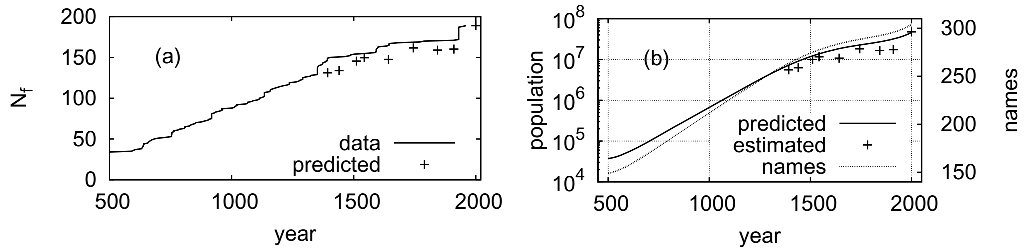

Figure 3. (a) The number of family names that are known to be introduced in Korea until a certain year (solid) and our predicted number of them (crosses), which is obtained by using the historical estimates (figure 2) and the family books in 1900–1990 (see text). (b) Our predicted size of population in the past (solid), based on the number of people in the same family books, carrying N_f different names shown in (a). The crosses are the same historical estimates as those in figure 2. Using this prediction as the input parameter of the RGF description (see (4)), we also estimate the *total* number of family names (dotted).

Korea increased by roughly a factor of 6, from about 8 to 46 mil. This increase is linear in time up to 1900 and then increases faster (compare figure 2). During the same period, the number of family names only increased very slowly, again with a sharp increase around 1900. In the rough estimate given below, the increase is about 27 names from around the year 1500 to 2000. In spite of this, the results shown in figure 1 suggest that this time evolution is such that $N(t)$ depends on time t only through $M(t)$, such that at all times $N(M)$ is a unique function. The uniqueness of the function $N(M)$ is also a consequence of the RGF model. In the context of word frequencies, it corresponds to the meta-book concept discussed in [6, 10]: the word frequency used by an author is to good approximation given by a unique function $N(M)$, where N is the number of distinct words and M is the total number of words, independent of which book, of what length book or when it was written. In short, figure 1 suggests that N is only a function of $M(t)$ and this particular feature is also consistent with the RGF model.

In order to illustrate the consequences of a time-independent distribution further, we use the data from [5] for the introduction years of 189 families. Figure 3(a) shows the number of these 189 families that existed at a given time from the year 500 to 2000. Note that the increase is slow: from year 1500 to 2000 the increase is only about 20%. Now imagine that you pick a group of male persons that belong to one of these 189 families in the year 2000. If we follow the lineage of this group of m persons back in time, the only decrease in the number of family names is caused by the fact that the lineage stops. Suppose that you randomly pick a fraction m/M of the population for the year 2000, which on average contains 189 different names. According to the family books in 1900–1990 (table 1), this number amounts to $m \approx 1.4 \times 10^5$ (figure 1). Roughly half of this group is male and if we follow the lineage backward in time, the remaining lineage will be roughly proportional to the total population so that $m(t)/M(t) = \text{const}$. As the population decreases, the average number of names, n, is hence just given by $n(m)$ provided that this to a good approximation is a unique function independent of time. In figure 3(a), the crosses show the prediction from this assumption, using the information on the total population given in figure 2 and the estimated $N(M)$ given in figure 1. As seen here, both the sharp decrease from 2000 to 1900 and the slow decrease between 1900 and 1400 are correctly reproduced. In figure 3(b), we do the inverse, using the same assumption: the data for the 189 families given in figure 3(a) are used as an input to estimate the total population within the period 500–2000. The full curve in figure 3(b) gives the estimated population estimated by this method. A comparison with the historical estimates from [11] shows that the agreement is everywhere within a factor of 2 (see figure 3(b)). The estimate suggests that the population in about 500 AD was around 4×10^4 and exceeded 10^7 around 1400 AD. It should be noted that our estimate of the total population refers to all persons integrated in the society having a Korean family name. This might, of course, in previous times only be part of the actual population controlled by the society. Figure 3(b) also shows the expected number of family names based on the prediction for the total population possessing Korean family names. According to this estimate there were already around 150 family names around the year 500 AD. Note that the expected number also reflects the fact that family names can both be added and subtracted from the population in the course of history (some families simply go extinct). Thus the rate of increase for the total population is expected to be slower than for the 189 families which only include families that have survived until the year 2000. This observation is also consistent with our prediction: there are 155 names introduced during the period 500–2000 according to our data set plotted in figure 3(a), and our predicted increase of the total number during the same period also happens to be 155.

A second feature of the RGF model is that the largest group is always proportional to the total size of the data set [6]. In the context of Korean family names, this implies that the proportion of persons named ‘Kim’ in a randomly picked group of Koreans should be constant, irrespective of the historical time or size of the group. In order to test this, we again start from the three latest historical windows in table 1, i.e. 1900–1930, 1930–1960 and 1960–1990. Note that the period 1930–1960 is the largest group, so that M actually decreases with time during some period within 1930–1990. These three data points are plotted in figure 4 (open circles) and the straight line in the figure is obtained by the least square fitting to these three data points. This linear prediction, based on the data from 1900–1990, is then directly compared with the data for the 13 time windows from 1510 to 1900. The agreement is striking. In addition, the number of Kims from the census of the year 2000 is given by the asterisk. Thus the figure spans group sizes from $M = 33$ to $M = 4.6 \times 10^7$ and history from the year 1510 to 2000. From this perspective, the proportionality is borne out to an amazing degree. Is this a surprising result? All it is saying is that the total number of persons named Kim grows and decreases at precisely the same rate as the total population. If the number of persons belonging to *any* family grows and decreases precisely as the total population, then the result will follow. However, this is not the case. The ten families in table 1 show a rapid increase of inflow of women by marriage from 1510 to 1900. The number of inflowing women is an approximate measure of the total number of persons who carry one of the ten names. Figure 5 shows this rapid growth: a factor of about

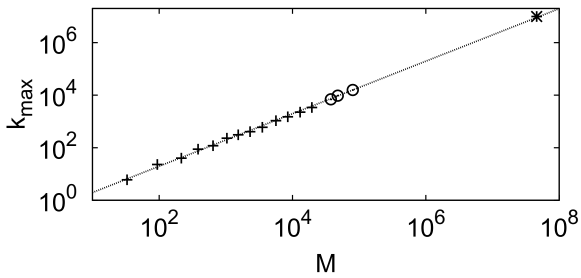

Figure 4. The number of persons with the most frequent family name in each family book. The circles on the upper right show the three most recent data sets of table 1, from which the slope a of the line has been determined by linear fitting, $k_{\max} = aM$. Note that these three points are not in time order since the middle point is the latest. The crosses are the remaining 13 time windows for which time and size are in the same order. The asterisk is the number of Kims for the whole population according to the census in the year 2000. The proportionality with the size of the group and the number of Kims is borne out to a very high degree irrespective of time.

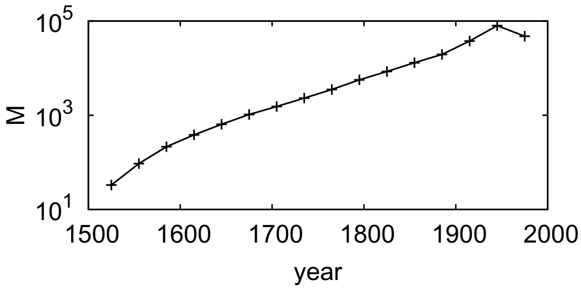

Figure 5. M as a function of time in table 1.

60 to be compared with a factor of about 2 for the total population during the same period. Thus individual name groups, in general, grow and decrease in a very different way compared with the total population. Nor do growth models, in general, predict that the largest group grows linearly with the size. For example, the Simon model predicts that the largest group grows like $k_{\max} \propto M^{1-\alpha}$, where $0 < \alpha < 1$ is the probability for a new name appearing during a time step [9, 12]. This means that only in the trivial case when the whole population is named Kim (which corresponds to $\alpha = 0$) is there a direct proportionality. Thus, in spite of the simplicity of the result, it appears to be nontrivial. Nevertheless, it is consistent with the prediction of the RGF model. We also note that combining our population estimate of persons with Korean family names with the size proportionality of Kim suggests that in around 500 AD there were already about 10 000 Koreans named Kim.

So far, we have only discussed two features of the RGF model: the uniqueness and time independence of the function $N(M)$ and the proportionality between the size of the largest group for any random part of a population. However, the RGF model also gives a precise prediction for the actual group-size distribution $P_M(k)$ in the form of (1). In order to test this prediction, we again start from the data in table 1 and the period 1900–1990 which contains

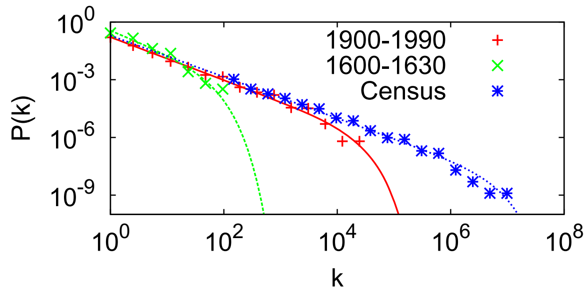

Figure 6. Comparison between the actual family books (points) and RGF predictions (lines).

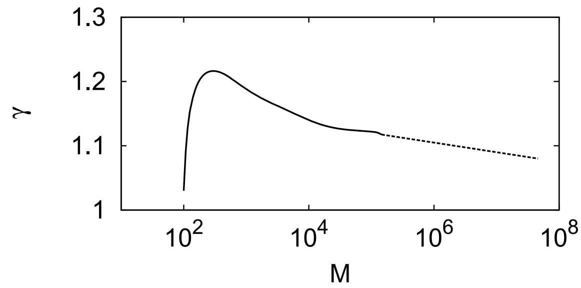

Figure 7. Power-law exponent as a function of M. The solid line was obtained by analyzing the family books from 1900–1990, and the dotted line is connected with the census data in 2000.

$M_{\mathrm{tot}} = 165\,020$ women each having one family name out of $N_{\mathrm{tot}} = 194$ and where $k_{\mathrm{max}} = 32\,316$ are named Kim. These three numbers M_{tot}, N_{tot} and k_{max} uniquely determine $P_M(k)$ within the RGF model, as explained in section 3 [6]. The middle full curve in figure 6 gives the predicted size distribution and the pluses denote the actual (binned) data points. The agreement between the RGF prediction and the data is very good, in particular in view of the fact that the prediction is based solely on the three numbers M_{tot}, N_{tot} and k_{max}. The prediction for the exponent γ in (1) is $\gamma = 1.12$. As explained in section 3, the RGF model allows you to predict how the $P_M(k)$ for either a smaller $m < M$ or larger $m > M$. Figure 7 displays the predicted change of γ when starting from the data given for the period 1900–1990. The solid curve gives the change when m is decreasing and the dotted line when it is increasing. The fact that γ changes with the size of the data set is a fundamental feature of the RGF model and distinguishes it from the usual growth models, which in general give scale-invariant and hence size-independent γ [6]. The left full curve in figure 6 is the prediction for the 1600–1630 data only using the data for 1900–1990 and the number $m = 384$, which is the number of women getting married into the ten families during the period 1600–1630. The actual name-frequency distribution for these women is given by the crosses and the agreement is again quite good. In particular, note that the data are indeed consistent with the slightly steeper slope for smaller k caused by a slightly larger $\gamma = 1.22$ (compare figure 7). The rightmost curve in figure 6, in the same way, gives the prediction based on only using the data for the married women in 1900–1990 and the total population size in the year 2000 given by $m = 4.6 \times 10^7$. The census data from the year 2000 are also plotted and the agreement between the prediction and the data is again very good. This time the $\gamma = 1.07$ is even closer to one. In short, the RGF model gives very good predictions for the distribution of actual Korean name-group sizes both backward and forward in time.

5. Concluding remarks

Our analysis suggests that the family-name distribution within the Korean population shares characteristic features with the word-frequency distribution for an author: both are to a good approximation described by a 'meta-book' distribution $N(M)$ and both are well described by the RGF model [6, 10]. The 'meta-book' distribution for an author gives a unique relation between a text of length M written by an author and the number of distinct words used, N. In the Korean case, this corresponds to the number of distinct family names N you typically find in a group of M Koreans. In the word-frequency case, this leads to the conclusion that the most common word used by an author in an English text, *the*, is proportional to the total size of the text M. The corresponding conclusion for the Korean family names is that the most common name, *Kim*, in a group of M Koreans should on average always be proportional to the size M. This prediction was checked with data ranging from 1510 AD to 2000 AD and group sizes in the interval $[33, 4.6 \times 10^7]$ and was found to be obeyed with high precision. It was argued that this is a nontrivial result for two reasons: firstly, it was shown that the rise and fall of individual families, in general, have no simple relation to the rise and fall of the total population; the only obvious relation is that the members of all the families collectively varies as the total population. Secondly, usual growth models, such as the Simon model, predict that the size of the largest family grows more slowly than the population.

The fact that the name distribution to a good approximation appears to follow a unique $N(M)$ made it possible to estimate the size of the Korean culture down to the year 500 AD by using the statistics for the years in which 189 Korean family names were introduced. This estimate suggested that the total population was around 5×10^4 persons with Korean family names of which about 10^4 carried the family name Kim. The total number of family names in the year 500 AD was predicted to be around 150. We believe that these are fascinating conclusions, although we cannot judge the historical realism and implications of the ten thousand Kims.

Finally, we demonstrated that in the actual frequency distributions, Korean names follow the RGF distribution with a size-dependent power-law exponent γ [6].

What do these results imply for the Korean culture? We speculate that the answer is stability. It seems that some core of the Korean culture has remained intact over at least 1500 years and as both the population and occupied area expanded, it basically swallowed other cultural influences without compromising its core. An interesting question is whether this type of analysis could also be applied to other cultures. This, however, remains to be investigated in the future.

Acknowledgments

SKB and PM acknowledge support from the Swedish Research Council through grant no. 621-2008-4449 and BJK from the Basic Science Research Program of the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology (2010-0008758).

References

[1] Miyazima S, Lee Y, Nagamine T and Miyajima H 2000 Power-law distribution of family names in japanese societies *Physica* A **278** 282

[2] Zanette D H and Manrubia S C 2001 Vertical transmission of culture and the distribution of family names *Physica* A **295** 1

[3] Reed W J and Hughes B D 2003 On the distribution of family names *Physica* A **319** 579

[4] Kim B J and Park S M 2005 Distribution of Korean family names *Physica* A **347** 683

[5] Baek S K, Kiet H A T and Kim B J 2007 Family name distributions: master equation approach *Phys. Rev.* E **76** 046113

[6] Baek S K, Bernhardsson S and Minnhagen P 2011 Zipf's law unzipped *New J. Phys.* **13** 043004

[7] Kiet H A T, Baek S K, Jeong H and Kim B J 2007 Korean family name distribution in the past *J. Korean Phys. Soc.* **51** 1812

[8] Cover T M and Thomas J A 2006 *Elements of Information* (New Work: Wiley)

[9] Bernhardsson S, Rocha L E C and Minnhagen P 2010 Size dependent word frequencies and translational invariance of books *Physica* A **389** 330

[10] Bernhardsson S, Rocha L E C and Minnhagen P 2009 The meta book and size-dependent properties of written language *New J. Phys.* **11** 123015

[11] Ko D-H (The Organization of Korean Historians) 2005 *How People Lived in the Joseon Era* 2nd edn (Cheongnyeonsa, Paju) chapter 1, p 19 (in Korean)

[12] Simon H A 1955 On a class of skew distribution functions *Biometrika* **42** 425